Relationships between quantitative variables

Statistics for Social Scientists I, L6

Bur, GJM

2024-07-01

1 Relationships between variables: prediction vs. simple association

Let's now begin discussing relationships between variables. We are starting **bivariate analysis**. We have previously *alluded* to this, mostly by doing *prediction*-type operations, e.g. working out the average number of children someone has, conditional on whether or not they are married. We'll show later that this is actually the best possible prediction[1]—so we have already been doing prediction!

It turns out that in the case of two quantitative variable, we *also* have a more-general but useful type of bivariate analysis which is not *predictive*, but instead quantifies the relationship more abstractly. It is much easier to talk about prediction between quantitative variables if we first review this measure of association.

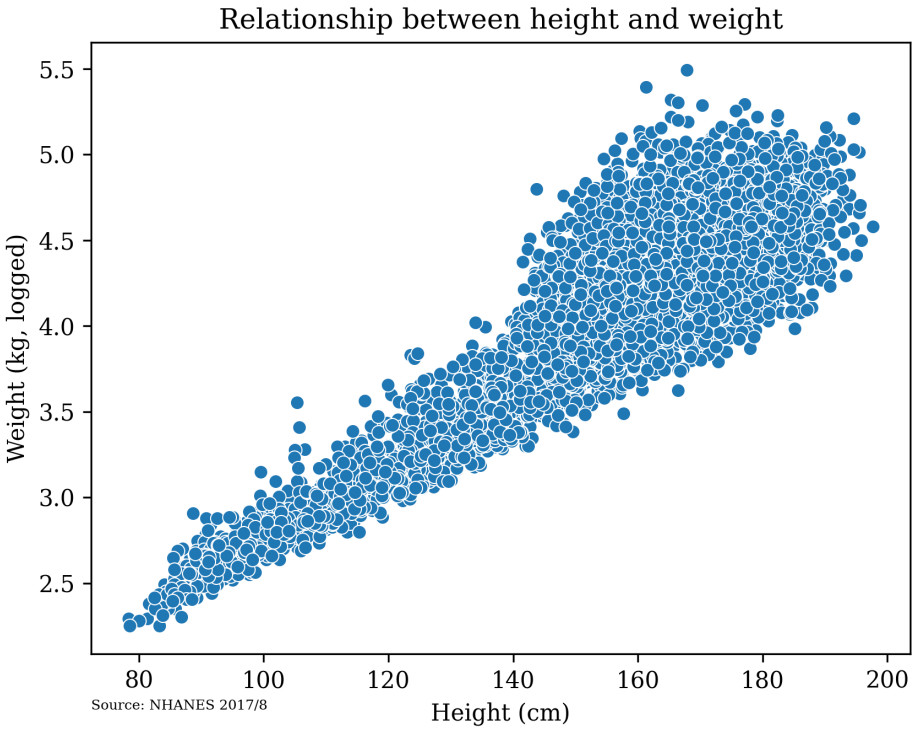
Let's spoil the ending: this measure of association is the correlation coefficient. Below, I motivate the definition of the correlation coefficient.

2 Correlation: the big idea

How do we measure the association between two variables? If I have a *scatter-plot*, how do I quantify the variables' relationship? Let's try to do so with *words* to start. Examine the plot below. How strong does the relationship feel? Does it seem linear or not?[2] Is it positive or negative? Do you spot any potentially high-influence observations?

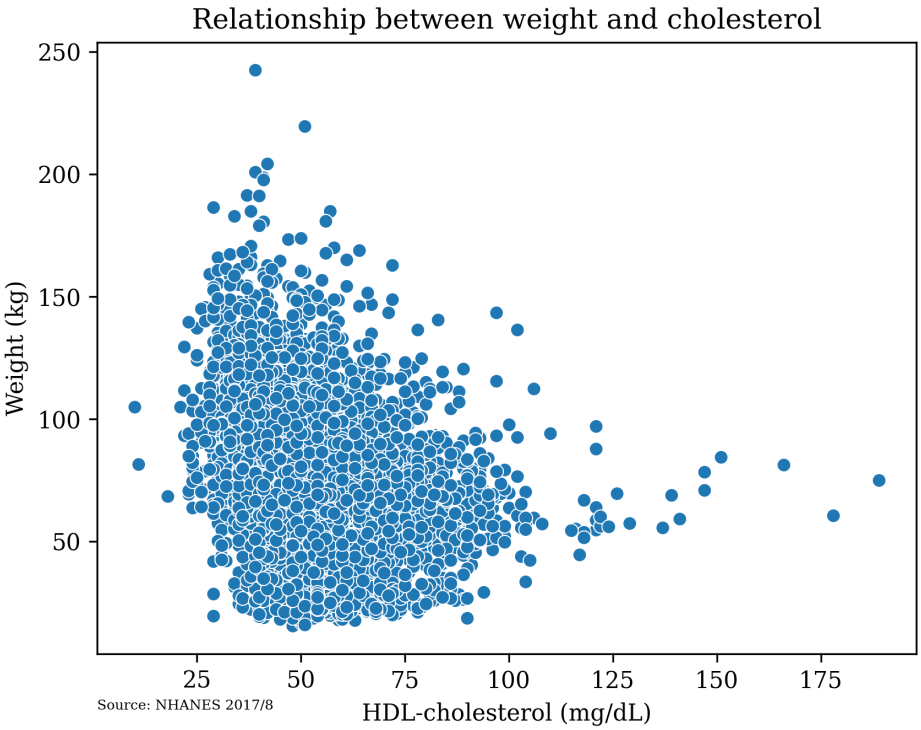
[1] Assuming that we use the standard loss function, i.e. the total or mean squared error. That's the only one we use in this course; it is by far the most common one in use.

[2] Ignore the logarithm for now. It matters for substantially interpreting the underlying variables, weight and height, but just pretend "log(weight)" is its own thing for now.



So, this is pretty clearly a linear relationship. It's quite strong, I think, but that's subjective. And, it's positive, with no obvious outliers.

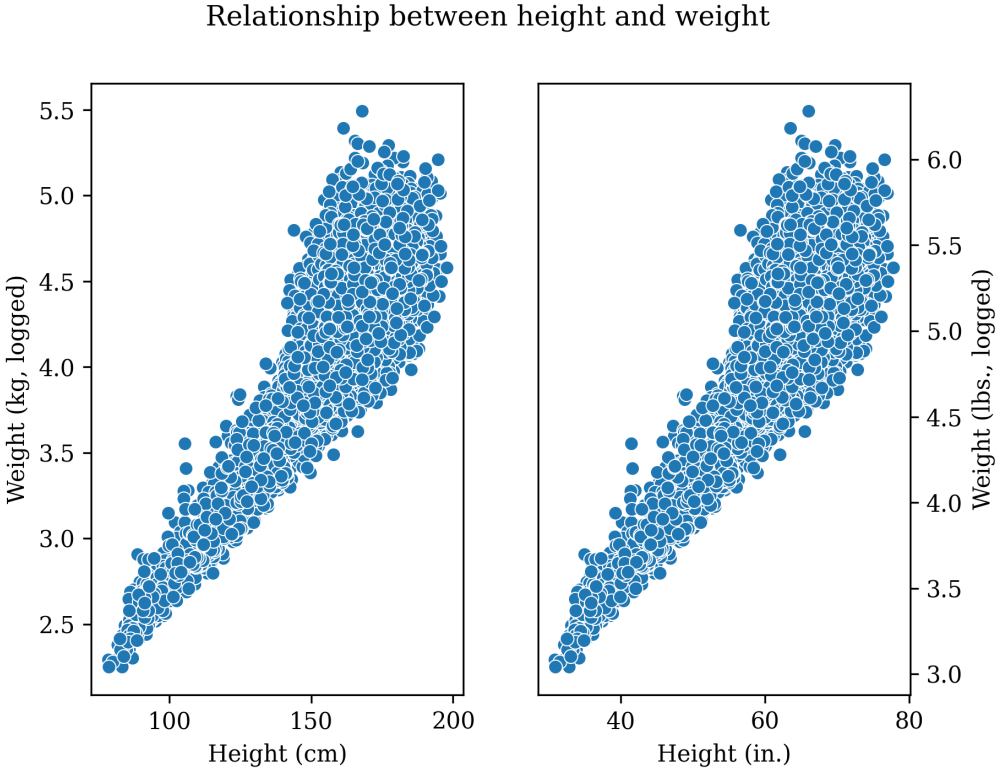
Let's try one more, after which point, we'll get into the heavier stuff.



The relationship here is much less strong, and much less linear. We do have some outliers that probably would make a naïve measure of association quite large. Outliers or note, the relationship is strikingly *negative*, an irritating fact for people who believe cholesterol is simply *bad* (as opposed to playing a very complex role in metabolism, but enough about my unconventional views on fitness).

So, how could we go about calculating such a measure of association? I'll show two ways into the problem.

First, what if we want to ensure that the relationship is *invariant* to scaling? For example, if I change our units on the variables above to American-friendly British Imperial measurements, the scatterplots look like this. How much should our measure of association change? You'll probably say "not at all"!



Secondly, we might want for our measure to have some kind of comparability across cases and to indicate *direction*. For example, the relationship between logged weight and height is clearly stronger than that between weight and cholesterol. It is also reversed. So, we might want a measure that also varies in a known range, say, $-1 \leq r \leq 1$, where r is our sample measure of **r**elation.

3 Correlation: a geometric derivation

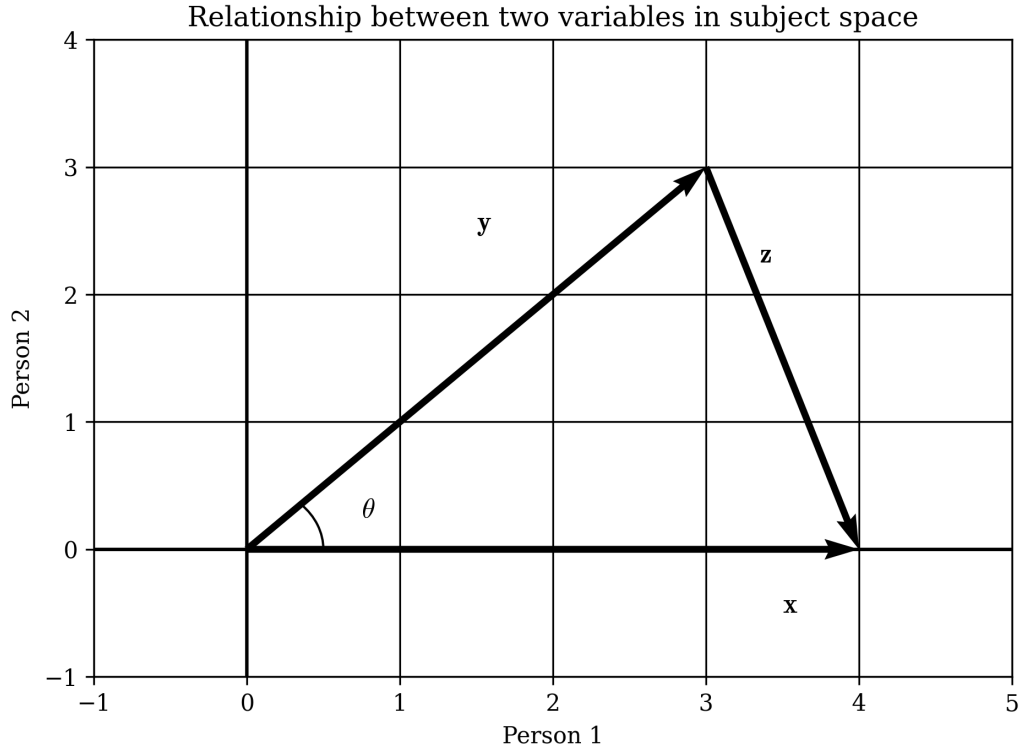
It turns out that the measure of association which has these properties is the correlation coefficient. It is written ρ in the population and r in the sample.[3] For example, in our examples above, our correlation coefficients were, respectively, $r = 0.89, -0.17$. **The strength is measured by the absolute value (relative to 1), and the direction by the sign.**

Before I show you the formula, I want to motivate this geometrically, but in a different way. Recall that above, we were working in what is called *variable space*: the axes of that two dimensional space each represent a variable. But, recall that we also can picture our variables in what is called *subject space*, where each axis is a person. Then, variables are represented by vectors, not axes: a vector is both a geometric object (just a point, really)

[3] As with μ, the letter ρ unfortunately gave rise to *one* Roman letter (m and r, respectively) but came to resemble a different Roman letter over thousands of years of changes in typesetting (u and p, respectively). Don't worry too much about this. The letters are meant to be intuitive and to remind you of what they stand for.

and also a list of coordinates, so it is a natural way to think about a list of scores like we have in statistics. Recall that this is an especially natural tool if we think about *centering* the original scores; so, for example, the whole vector of scores $\mathbf{y}_c = \mathbf{y} - \bar{\mathbf{y}}$.

Here is a picture of how that looks in just two dimensions. On the variable x person 1 had a (centered) score of 4; person 2 had a score of zero (centered; I won't write that explicitly anymore). So, that variable is the vector $(4, 0)$ (a vector is just a point drawn as an arrow). On variable y, person 1 had a 3, while person 2 had a 3 also, so that is the vector $(3, 3)$. The side labeled $\mathbf{z}$ is just the difference between the two vectors, which we won't need directly, but it is useful in understanding the proof I give in the appendix.



Now, in this subject space, we have a pretty obvious measure of association, more obvious than with the scatterplot. That measure of association is in fact the angle between the vectors! However, directly calculating this angle is a bit tricky.

Fortunately, the *cosine* of the angle between the vectors has the desired properties! First, $-1 \leq \cos(\theta) \leq 1$. Second, the actual formula for finding the cosine will resist changes to either variable that are *superficial* from the standpoint of relationship between them.

Finally, *even better*, the the law of cosines (proof in appendix) tells us that for all triangles, not just right triangles, there is a somewhat-scary looking formula that actually involves many quantities with which we've already dealt, making it rather simple.

So, the correlation coefficient is simply *defined to be* the cosine of the angle between two vectors that represent *centered* variables. Now, how do we actually calculate it?

4 The correlation formula: useful alternatives

The law of cosines tells us that to find this cosine of this angle in n-dimensional space, we write…

$$\cos(\theta_{x,y}) = r_{x,y} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2} \cdot \sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}}$$

There are a couple of very useful rewrites of this formula.

First, if we add an $n-1$ to the denominators of the numerator and the denominator (check this if you doubt me; it requires a bit of algebra and remembering root rules), we get some familiar quantities…

$$r_{x,y} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{(n-1)s_x s_y}$$

The quantity $\frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{(n-1)}$ is itself known as the **covariance** and written $s^2_{x,y}$. It is just the unstandardized, easier-to-compute-but-less-intuitive version of the correlation, much like the variance is to the standard deviation. But, it comes up often in derivations as an intermediary quantity, so we give it a name. So, we can write our formula like so: $\frac{s^2_{x,y}}{s_x s_y}$. Or, with lighter notation, we have "the covariance of x and y divided by the product of their standard deviations".

Second, you might notice that this all of this kind of resembles the z-scores we were computing before! Again, squint a little bit: if $n - 1 \approx n$, we have a mean of the pointwise product of z-scores! By the way, the convenient term for the "sum of a pointwise product of two variables" is the **dot product**. So we can just say that correlation is the **averaged dot-product of standardized variables**.

$$\begin{aligned} r_{x,y} &= \frac{1}{n-1}\sum_{j=1}^{n}\left\{\frac{(x_j - \bar{x})}{s_x}\frac{(y_j - \bar{y})}{s_y}\right\} \\ &= \frac{1}{n-1}\sum_{j=1}^{n} z_{xj} z_{yj} \\ &\approx \overline{z_x z_y} \end{aligned}$$

Yet one more, you ask? We also have a close cousin of the K-H identity here. I'll move the proof to the appendix, but the following is true.

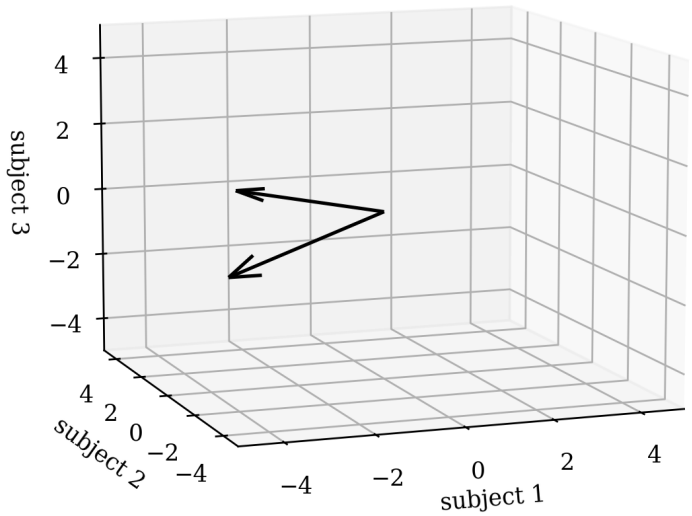
$$r_{x,y} \approx \frac{1}{s_y s_x} \overline{xy} - \bar{x}\bar{y}$$

In words, the correlation is *roughly* equal in the sample (and exactly equal in the population) to the mean of the product less the product of the means, all divided by the product of standard deviations. It's almost incredible how neatly this all works out!

5 Correlation: computational practice

Let's get a bit more practice. First, why don't we revisit our two variables that we previously visualized in subject space. Below the picture are the data. Can you find the correlation coefficient? I went ahead and calculated the mean for you already and then the deviations.

Variance in subject-space



$\mathbf{j}$	$\mathbf{y_j}$	$\mathbf{y_j - \overline{y}}$	$\mathbf{x_j}$	$\mathbf{x_j - \overline{x}}$
1	3	-2	4	-2.67
2	10	5	9	2.33
3	2	-3	7	0.33

Our task is now pretty simple: first, the root sums of squares are $\sqrt{(-2)^2 + 5^2 + (-3)^2} =$

6.16 and $\sqrt{(-2.67)^2 + (2.33)^2 + (0.33)^2} = 3.56$. Then, the sum of the products of the centered variables is $(-2)(-2.67) + (5 \cdot 2.33) - (3 \cdot 0.33) = 16$. Then, $\frac{16}{6.16 \cdot 3.56} = 0.73$.

Another way to do this is as follows. First, find the standard deviation and mean for each group. Here, $\bar{y} = 5, \bar{x} = 6.67$. We already found the sums of squares for each group (and sometimes this information will just be given), so we can use those to find the standard deviations quickly: $s_y = \frac{6.16}{\sqrt{2}} = 4.36; s_x = \frac{3.46}{\sqrt{2}} = 2.45$ (our rounding error here is a bit worse than usual thanks to sample size, but don't worry much).

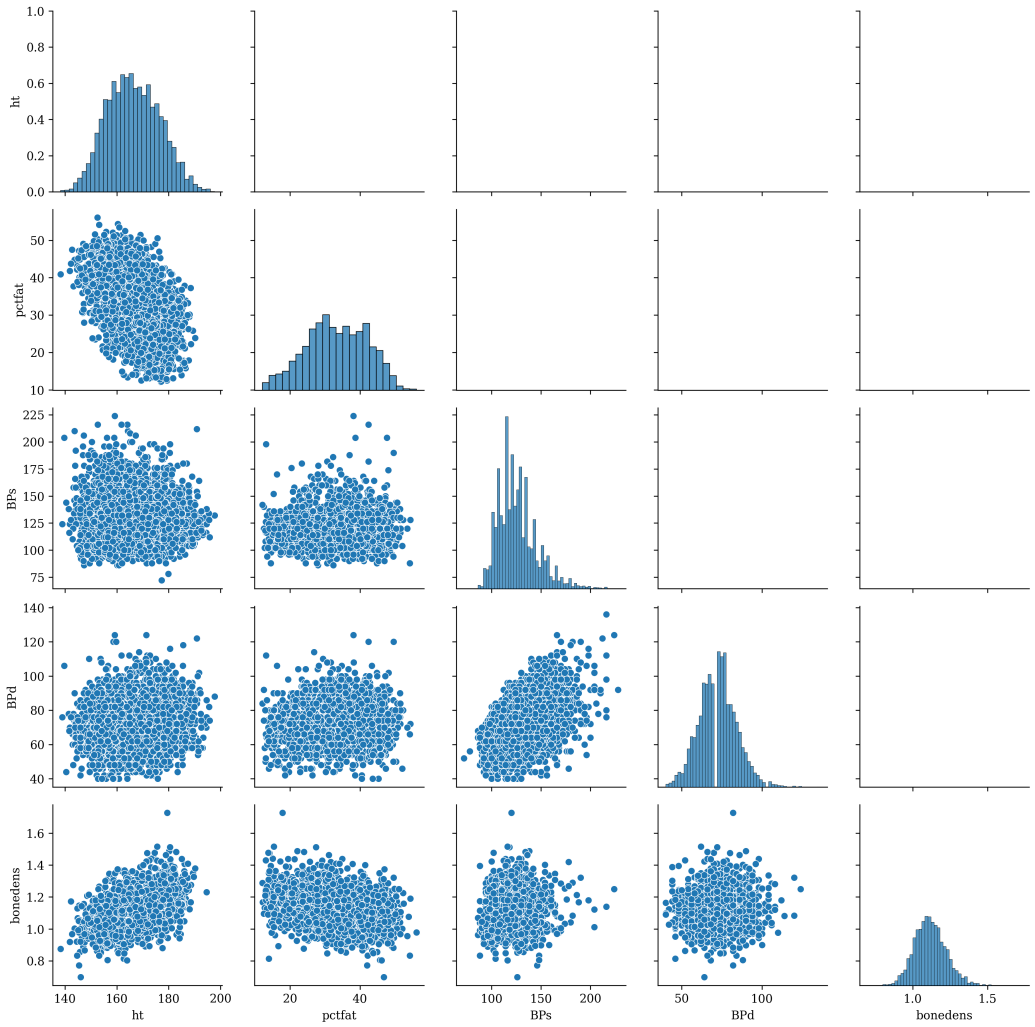
Now, let's find z-scores for each observation. You might want to put this into a table at this point to facilitate the calculation. It might look something like this.

	x	y	x_z	y_z	z_prod
1	4.00	3.00	-1.06	-0.46	0.49
2	9.00	10.00	0.93	1.15	1.06
3	7.00	2.00	0.13	-0.69	-0.09
total	20.00	15.00	-0.00	0.00	1.46

Dividing the total z-product by $n - 1 = 2$, we get 0.73, the same as before.

6 Correlation: some graphical practice

Let's get a bit more practice relating scatterplots in variable space to the correlation coefficient. Here are some examples of scatterplots between body measurement variables drawn from the NHANES data. The variables are a respondent's height, percent body fat, systolic blood pressure, diastolic blood pressure, and bone density. Can you guess the correlations without looking?



Here are the correlations. Are you surprised by any, or were your guesses ≈ accurate? Why is this matrix (read: table of data) symmetric? Why does it have 1s on the main diagonal?

	ht	pctfat	BPs	BPd	bonedens
ht	1	-0.51219	0.105742	0.156847	0.424462
pctfat	-0.51219	1	0.00156916	0.0125039	-0.319286
BPs	0.105742	0.00156916	1	0.645934	0.114128
BPd	0.156847	0.0125039	0.645934	1	0.105293
bonedens	0.424462	-0.319286	0.114128	0.105293	1

7 Correlation: some caveats

So, correlation is pretty cool. Some caveats, though, as always.

1. **Correlation does not tell us anything about causality.** It just establishes associations.

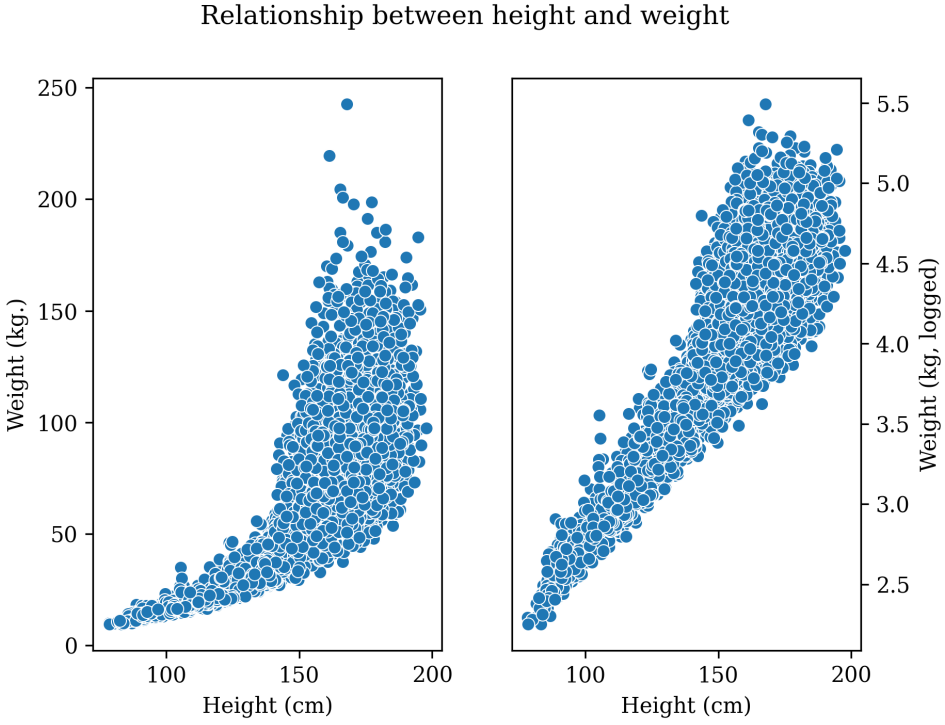
   In fact, in *all* of our examples above drawn from NHANES, causality is, at best, hard to infer.

2. **Correlation is symmetric with respect to the variables**. More prosaically, the correlation of X and Y is *one* thing. There are not two correlations for X, Y and Y, X. This won't always be true for comparable measures.

   You can just examine the formula and notice that there is no spot where order could really matter.

3. **Correlation does not tell us much about non-linear relationships**. It (wrongly) suggests that symmetric-ish non-linear relationships are zero; it tends to understand monotonic non-linear relationships and gives no clue about the non-linearity.

   For example, in the figure below, the correlation between height and weight at left is strong, at $r = 0.77$, but this does not tell us anything about the (obvious) non-linear pattern here, and I can make the correlation extremely strong, to a degree almost unknown in human statistics, by simply taking the logarithm of weight (figure at right, $r = 0.92$). The correlation coefficient is "silent" about this—you should always plot your data and examine these types of relationships.[4]

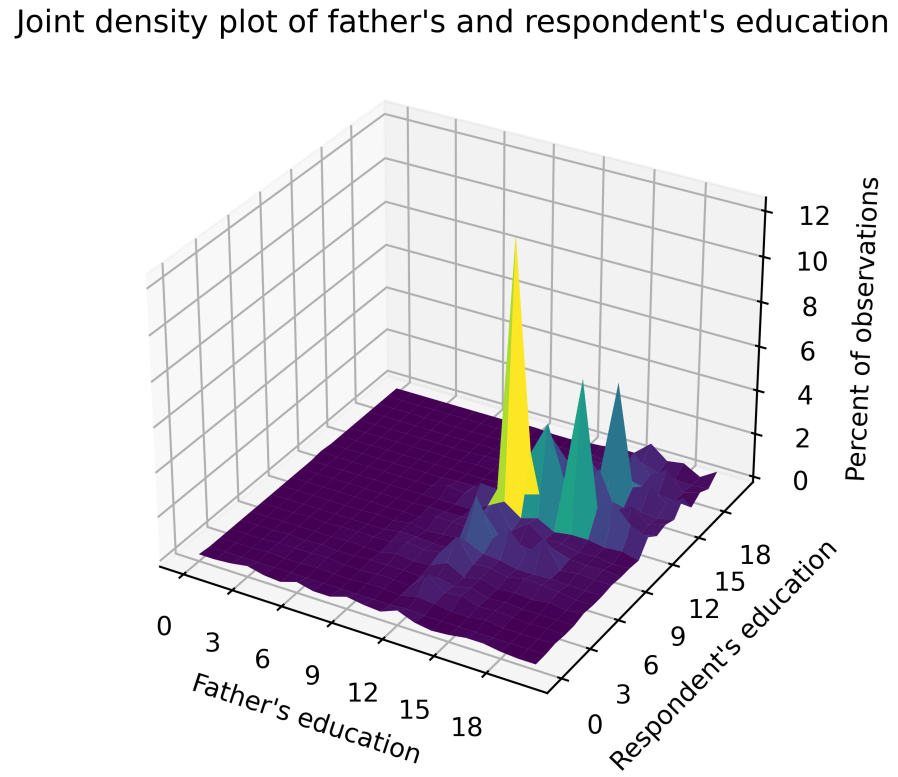


4. **Graphing tip: beware highly-discrete data**. Use a contour plot or a jitter.

[4] In fact, if you have data which are in a *perfect* quadratic relationship centered at the origin as in $y = x^2$, your correlation is zero.

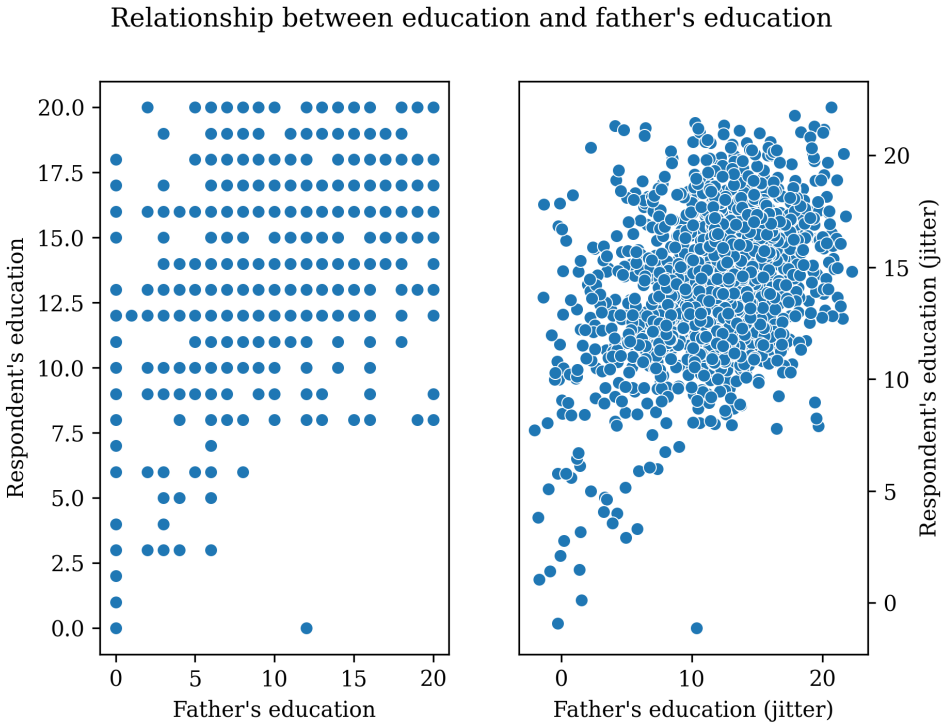
Scatterplots are technically what are called *joint* probability density functions (PDFs), analogous to the single-variable PDFs we saw before. If your data are highly-discrete, there will be a "pile-up" of density that is not visible on an ordinary scatterplot around certain key values.

That is, what we are really trying to visualize is this…



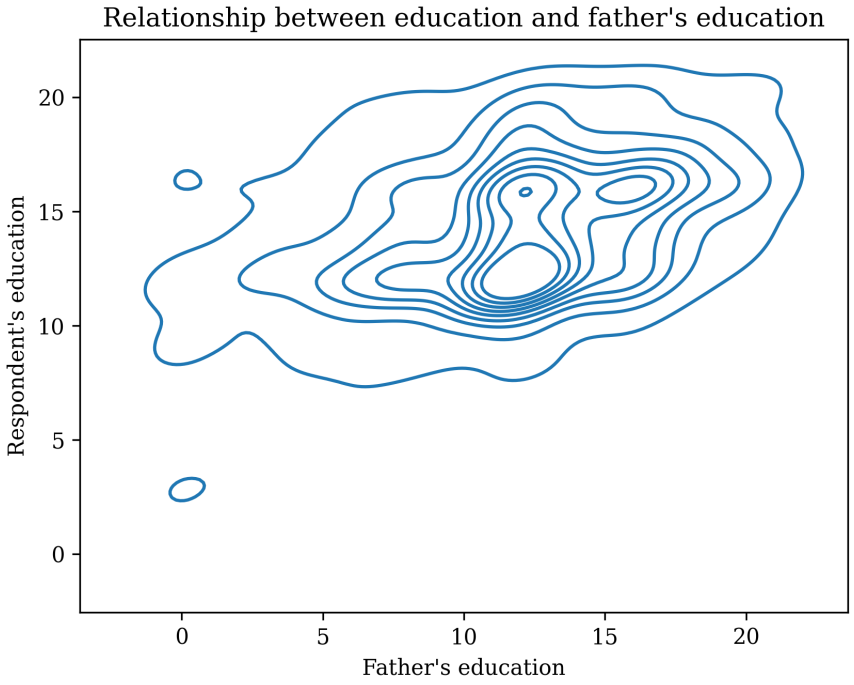
Source: GSS 2018, n=1687

The syntax to fix this in Stata is `scatter $y$ $x$, jitter($percent$)`. Below I fix it manually in Python by adding small Gaussian noise to the data.



The other main alternative is a contour plot. This produces level sets (two-dimensional curves which represent regions of the input space where the output, the joint density, is constant), using the same kernel density technique that we saw before. Again, prosaically, this is basically a map of the mountains. In Stata, try the following...

```
ssc install kdens2 // this is a community-contributed command
kdens2 $response$ $predictor$
```



8 Exercises

0. Set up a do-file with the standard header.

```
DATE: 2024-07-$dd$
AUTHOR: $LASTNAME, FIRST INITIAL$
TASK: Learn about correlation

capture log close
cd $/path/to/your/folder$
log using "2024-07-$dd$-L5-$LASTNAME$-$FIRSTINITIAL$", text replace
```

1. Let's use another random subset of the Titanic data. Many people had trouble with the CSV version of the data, so I put up a Stata-formatted data-set to the course Github that you can just download and directly do. For the brave, you can try the first line below. Just remember to un-comment it. You might need to rename some variables still.

```
* import delimited using "https://tinyurl.com/soc360su24titanic", varnames(12) clear
* If that command doesn't properly load in the data with names,
* just go visit https://github.com/griffinjmbur/soc360su24public/tree/mainSU24/Week3
* and download the data, then put it into your folder for the class
* and write "use titanic"
rename *, lower
```

```
list age fare if inlist(passengerid, 591, 132, 629, 196, 231)
br
```

2. Find the correlation coefficient by hand between `age` and `fare` for the five people involved. You're welcome to use any method that you like. There are two that are generally useful. First, you can make z-scores for each individual on each variable, then take the product for each person $z_{xj}z_{yj}$, then sum and divide by $n-1$. Or, you can make centered columns as shown above, using this to find the root sums of squares and the summed product of the centered variables. Then, just divide the latter by the former.

   I'll give you the table of summary statistics to save you some time using whichever approach you want.

	age	fare
mean	34.8	50.4133
std	14.4465	63.0312

3. Now, let's actually let Stata do some of our work. Calculate the correlation coefficient first only for our small set, then for all the data. What do you notice? Was our random subset representative of the data or not?

   ```
   corr age fare if inlist(passengerid, 591, 132, 629, 196, 231)
   corr age fare
   ```

4. Make a scatter plot of our data (`scatter age fare`). Do these data seem a bit jumbled? How can we spread them out a bit without radically changing them?

9 Exercise answers

1. Shown above.
2. Your table with centered variables should resemble this.

	fare	fare_c	age	age_c
590	7.12	-43.29	35.00	0.20
131	7.05	-43.36	20.00	-14.80
628	7.90	-42.52	26.00	-8.80
195	146.52	96.11	58.00	23.20
230	83.47	33.06	35.00	0.20

   Your table with the squares and the cross-product in it should resemble this.

	age	age_c	age_c_sq	fare	fare_c	fare_c_sq	fare_cxage_c
590	35.00	0.20	0.04	7.12	-43.29	1873.88	-8.66
131	20.00	-14.80	219.04	7.05	-43.36	1880.38	641.78

	age	age_c	age_c_sq	fare	fare_c	fare_c_sq	fare_cxage_c
628	26.00	-8.80	77.44	7.90	-42.52	1807.74	374.15
195	58.00	23.20	538.24	146.52	96.11	9236.65	2229.69
230	35.00	0.20	0.04	83.47	33.06	1093.07	6.61

And with sums it should resemble this. Note the expected behavior that the sums of the centered columns are zero (you didn't need to actually calculate this; Python just did it for me automatically when I wrote the script).

	fare	fare_c	age	age_c	fare_c_sq	age_c_sq	fare_cxage_c
590	7.12	-43.29	35.00	0.20	1873.88	0.04	-8.66
131	7.05	-43.36	20.00	-14.80	1880.38	219.04	641.78
628	7.90	-42.52	26.00	-8.80	1807.74	77.44	374.15
195	146.52	96.11	58.00	23.20	9236.65	538.24	2229.69
230	83.47	33.06	35.00	0.20	1093.07	0.04	6.61
sum	252.07	0.00	174.00	0.00	15891.72	834.80	3243.58

Finally, dividing, we have 3243.58 divided by the product $\sqrt{15891.72 \cdot 834.80} = 0.89$.

An alternate method that I ask you to use in the homework is the "(near)-mean dot product of z-scores" approach. You can craft a table of z-scores that looks like this.

	age	age_z	fare	fare_z	zprod
590	35.00	0.01	7.12	-0.69	-0.01
131	20.00	-1.02	7.05	-0.69	0.70
628	26.00	-0.61	7.90	-0.67	0.41
195	58.00	1.61	146.52	1.52	2.45
230	35.00	0.01	83.47	0.52	0.01
total	174.00	0.00	252.07	-0.00	3.56

Then, we simply take the `total` row of the `zprod` column (this is the dot-product of the z-scores) and take the pseudo-mean, dividing by $n-1 = 4$. We get $r = \frac{3.56}{4} = 0.89$.

3. The correlation for the whole data-set was very different at only about 0.1.
4. `scatter fare age, jitter(10) title("Relationship between age and fare on the {it:Titanic}")`

10 Appendix

10.1 The König-Huygens identity for correlation

Let's start with our original expression.

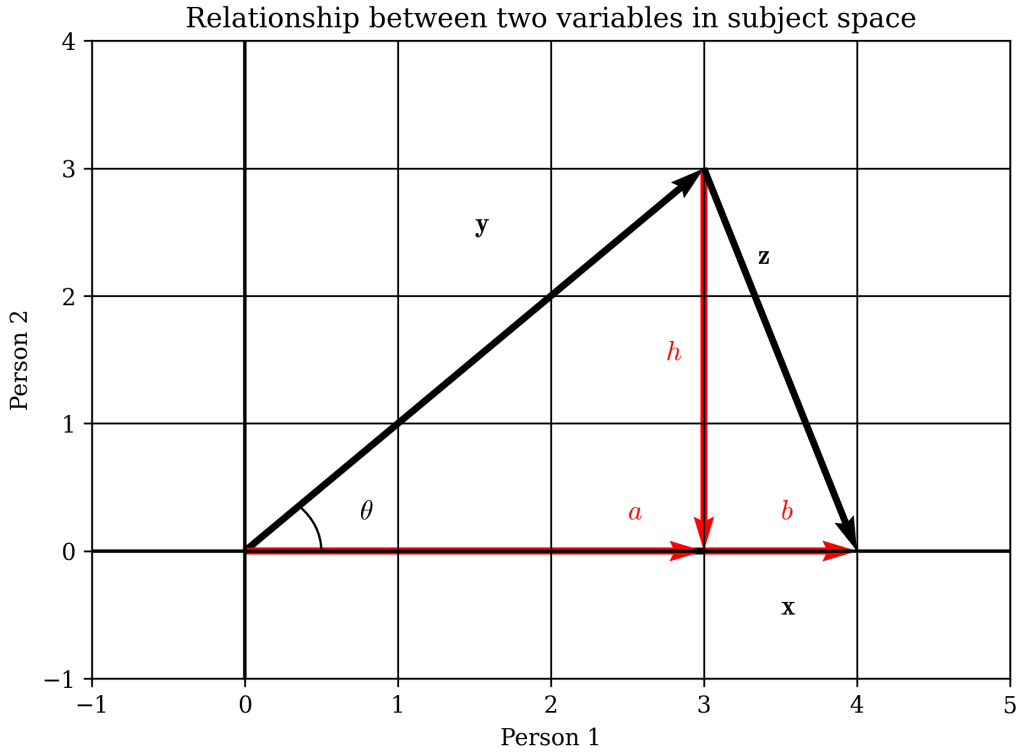
$$
\begin{aligned}
r_{x,y} &= \frac{1}{\sqrt{\sum_{j=1}(x_j - \bar{x})} \cdot \sqrt{\sum_{j=1}(y_j - \bar{y})^2}} \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) \\
&= \frac{1}{\sqrt{\sum_{j=1}(x_j - \bar{x})} \cdot \sqrt{\sum_{j=1}(y_j - \bar{y})^2}} \sum_{j=1}^{n}(x_j y_j - \bar{x} y_j - \bar{y}\bar{x} - \bar{y} x_j) \\
&= \frac{1}{\sqrt{\sum_{j=1}(x_j - \bar{x})} \cdot \sqrt{\sum_{j=1}(y_j - \bar{y})^2}} \sum_{j=1}^{n}(x_j y_j) - \bar{x}\sum_{j=1}^{n} y_j - \bar{y}\sum_{j=1}^{n} x_j + n\bar{y}\bar{x} \\
&= \frac{1}{\sqrt{\sum_{j=1}(x_j - \bar{x})} \cdot \sqrt{\sum_{j=1}(y_j - \bar{y})^2}} \sum_{j=1}^{n}(x_j y_j) - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{y}\bar{x} \\
&= \frac{1}{\sqrt{\sum_{j=1}(x_j - \bar{x})} \cdot \sqrt{\sum_{j=1}(y_j - \bar{y})^2}} \sum_{j=1}^{n}(x_j y_j) - n\bar{x}\bar{y}
\end{aligned}
$$

Then, using the fact that we can stuff an $n-1$ beneath both numerator and denominator, we have…

$$
r_{x,y} \approx \frac{1}{s_y s_x}\overline{xy} - \bar{x}\bar{y}
$$

10.2 The law of cosines

The law of cosines can be proven by taking our original triangle above, dropping a perpendicular, and labeling the auxiliary parts of the triangle as shown below.



Then, we can use the regular Pythagorean theorem as follows.

$$\begin{aligned} b^2 + h^2 &= z^2 \\ (x-a)^2 + h^2 &= z^2 \\ x^2 + a^2 - 2xa + h^2 &= z^2 \\ x^2 + y^2 - 2xa &= z^2 \\ x^2 + y^2 - 2xy\cos(z) &= z^2 \end{aligned}$$

This last expression is the "regular" law of cosines, albeit in unusual notation—standard would be something like $a^2 + b^2 - 2ab\cos(c) = c^2$.

10.3 Generalizing the law of cosines to n-dimensional spaces

Now, we need to generalize this to higher-dimensional spaces. First, we should note that, to be precise, the quantities that we normally use in these equations are *scalars*: they are lengths or squared lengths. So, let's write that more explicitly now.

$$||x||^2 + ||y||^2 - 2||x|| \cdot ||y|| \cdot \cos(z) = ||z||^2$$

Let's now replace these with explicit formula for the lengths of vectors.

$$\sum_{j=1}^{n}(x_j)^2+\sum_{j=1}^{n}(y_j)^2-2\sqrt{\sum_{j=1}^{n}(x_j)^2}\sqrt{\sum_{j=1}^{n}(y_j)^2}\cdot\cos(z)=\sum_{j=1}^{n}(z_j)^2$$

10.4 Connecting the generalized law of cosines to variables in person space

Finally, the last part is to realize that $\mathbf{z}_j=\mathbf{y}_j-\mathbf{x}_j$. Then, we replace that in our formula. Let's handle just that part for now.

$$\begin{aligned}\sum_{j=1}^{n}(z_j)^2&=\sum_{j=1}^{n}(y_j-x_j)^2\\&=\sum_{j=1}^{n}(x_j)^2+\sum_{j=1}^{n}(y_j)^2-2\sum_{j=1}^{n}(y_jx_j)\end{aligned}$$

We now see that much of our equation cancels.

$$\begin{aligned}\sum_{j=1}^{n}(x_j)^2+\sum_{j=1}^{n}(y_j)^2-2\sqrt{\sum_{j=1}^{n}(x_j)^2}\sqrt{\sum_{j=1}^{n}(y_j)^2}\cdot\cos(z)&=\sum_{j=1}^{n}(x_j)^2+\sum_{j=1}^{n}(y_j)^2-2\sum_{j=1}^{n}(y_jx_j)\\-2\sqrt{\sum_{j=1}^{n}(x_j)^2}\sqrt{\sum_{j=1}^{n}(y_j)^2}\cdot\cos(z)&=-2\sum_{j=1}^{n}(y_jx_j)\\\cos(z)&=\frac{\sum_{j=1}^{n}(y_jx_j)}{\sqrt{\sum_{j=1}^{n}(x_j)^2}\sqrt{\sum_{j=1}^{n}(y_j)^2}}\end{aligned}$$

Now, let's remember that we are actually considering our variables to be *centered* versions of the underlying variables of interest. So, let's see what happens when we plug in centered variables. I will also finally replace the *angle* z with $\theta_{x,y}$ since this notation is more meaningful.

$$\cos(\theta_{x,y})=\frac{\sum_{j=1}^{n}(y_j-\bar{y})(x_j-\bar{x})}{\sqrt{\sum_{j=1}^{n}(x_j-\bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_j-\bar{y})^2}}$$

Indeed, this is the correlation coefficient.